

Следует отметить, что Дальтон-план – это не метод, а рамки, формы организации работы учащихся, которые не определяют метод и в которые можно вложить какое угодно содержание. Значительно дальше в этом отношении идет система целевых актов или метод проектов.

Как только человеку предоставляется возможность принятия собственных решений, выбора пути, он становится спокойнее, организованнее, строже по отношению к себе. Он рассуждает, анализирует ситуацию, взвешивает возможности, выбирает варианты. Это начало саморазвития, самосозидания. Задача преподавателя, мастера при таком обучении – наблюдать за ходом работы, помогать каждому учащемуся самостоятельно достигать цели. Важным достижением развивающих систем обучения является развитие самосознания, осознание своей человеческой неповторимости, своих пределов и возможностей, формирование навыков самоконтроля, сознательной жизненной ориентации.

Литература

1. Мэй Р. Искусство психологического консультирования. М., 1994.
2. Штейнер Р. Методика обучения и предпосылки воспитания. М., 1994.

Л.В. Колясникова

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ТЕХНОЛОГИИ ТЕСТОВОГО КОНТРОЛЯ

С тех пор как существует обучение, осуществляется и контроль уровня усвоения учебного материала. Но традиционные способы оценки, используемые в нашей системе образования, нуждаются в систематическом дополнении объективными методами, и необходимо многое сделать для того, чтобы найти нужное обоснование методики оценок. С одной стороны, не следует сводить педагогическую диагностику к простому тестированию, а с другой – без разумного использования информативных тес-

тов невозможно добиться существенного улучшения оценочной практики.

Современная научно обоснованная дидактика должна опираться на богатый инструментарий максимально объективных методов педагогической диагностики. Этим инструментарием, созданным в течение последних ста лет, являются тесты. Что же такое тест? Тест (в современном понимании этого слова) связан в сознании людей с научно обоснованной процедурой, позволяющей выявить реальные интересующие нас качества объекта. Более точная формулировка звучит так: тест – это кратковременное, технически сравнительно простое испытание, проводимое в равных для всех испытуемых условиях и имеющее вид такого задания, решение которого поддается количественному учету и служит показателем степени развития к данному моменту известной функции у конкретного испытуемого.

Теперь постараемся ответить на вопрос, где и как необходимо использовать тестовые технологии, а также в чем заключается их преимущество перед другими формами контроля. Прежде всего, использование тестовых технологий позволило проводить массовые измерения.

1. *Абитуриентское тестирование.* В 1997 г. появилась централизованная ветвь тестирования абитуриентов (100 тыс. человек по России). После прохождения тестирования абитуриентам выдается сертификат. Тестирование проводится по русскому языку, биологии и физике. Нормы оценок используются более жесткие, чем в школе.

2. *Региональное тестирование* (независимый мониторинг результатов образования). Тестирование проводится в пунктах тестирования для выпускников 9-х классов. Результаты тестирования фиксируются в сертификате, который выдается учащемуся. С одной стороны, снижается нагрузка на школы, с другой стороны, если учащийся получает неудовлетворительный результат, он может не сообщать его в школу. В школы сообщают только обобщенные результаты.

3. *Аттестация общеобразовательных учреждений.* Каким образом оценить качество и уровень подготовки выпускников образовательных учреждений? Так как это массовая работа, то

контрольная работа или устные проверки не могут быть использованы. Оптимальным средством являются тесты.

Подчеркнем достоинства тестовых форм педагогического контроля.

1. Прежде всего, тестовый контроль является объективным методом диагностики успехов учащихся в обучении. Результат тестирования поддается однозначному и воспроизводимому измерению и оценке, так что педагог не может внести в него свои произвольные коррективы, как при субъективном контроле.

2. Тестовые методы оценивания используют процедуры, в которых роль человека, проводящего оценивание, минимальна. Она сводится к роли оператора. Кроме того, обработка результатов тестирования занимает минимум времени. Вследствие этого педагог в принципе может сосредоточиваться только на преподавании, больше времени уделять отдельному ученику, а не группе в целом.

3. Тесты – это *оптимизированные* испытания или средства педагогического контроля с целью количественного измерения качеств человека. Количественное измерение человеческих качеств (квалиметрия) опирается на мощные математические методы анализа результатов (теорию корреляций, регрессионный анализ, дисперсионный и факторный анализ, законы распределения, теорию вероятностей и математической статистики). Тестирование вводит в педагогику и другие науки о человеке процедуру измерения, которая позволяет описать свойство объекта наиболее точно. Если используемая повсеместно грубая пятибалльная шкала не позволяет ни ученику, ни его родителям отследить критический для учащегося переход от «двойки» к «тройке» или от «четверки» к «пятерке», то измерительные шкалы (такие как интервальная шкала), применяемые в тестовых технологиях, позволяют ответить на вопрос, насколько лучше (хуже) он стал учиться. Таким образом, шкалирование существенно облегчает процедуру сравнения при условии, что цифровые показатели шкал адекватно отражают действительность.

4. Использование тестовых технологий позволяет частично решить проблемы личности (в частности, осуществить выбор будущей профессии); образовательных учреждений (аттестация, обеспечение более высокого качества образования, сравнение ус-

пехов различных образовательных учреждений); органов управления образованием (для управления руководящими органами городов, округов, регионов необходимо создать строго научную, экономичную и хорошо отлаженную аттестационную систему, охватывающую все образовательные учреждения, находящиеся на данной территории).

Как и любое измерение, тест достигает цели только тогда, когда он является эффективным и качественным. В классической теории тестирования основными критериями качества теста являются надежность и валидность. Помимо них выделяют также дискриминативность и объективность тестов. Остановимся на них более подробно.

Надежность характеризует достоверность и точность результатов измерения знаний. Например, значение надежности 0,85 означает, что достоверность определения уровня знаний составляет 85%, а ошибка измерения – соответственно 15%. Педагогическое измерение считается состоявшимся, если надежность равна 0,8.

Различают надежность по внутренней согласованности и ретестовую надежность. Окончательный вариант теста должен соответствовать требованию внутренней согласованности. Это характеристика теста, указывающая на степень однородности состава заданий с точки зрения измеряемого качества (т. е. все задания теста должны определять тестируемое качество). Для ее определения применяется процедура установления корреляций между результатами теста в целом и каждым отдельным заданием. Коэффициент корреляции между вектором i -го задания и суммой баллов не должен быть меньше 0,3. Также тест считается надежным, если он даёт одни и те же результаты для каждого испытуемого при повторном тестировании. Такая надежность называется ретестовой. Оценка надежности основывается на классической теории погрешности измерений.

На практике используются три основных метода оценки надежности тестов:

- 1) повторное тестирование (ретестирование);
- 2) параллельное тестирование (тестирование с использованием параллельной формы теста);
- 3) расщепление (метод деления целого на части).

Основными достоинствами метода повторного тестирования являются естественность и простота определения коэффициента надежности, главным недостатком – неопределенность в выборе интервала между двумя тестированиями, в зависимости от чего по-разному проявляются такие факторы, как запоминание содержания задач и способов их решения, влияние опыта и общения испытуемых после первого тестирования.

Второй метод – определение надежности при тестировании с использованием параллельной формы теста. Параллельными считаются такие тесты, которые измеряют одно и то же свойство.

При применении параллельных форм теста испытуемые делятся на две группы (разделение проводят случайным образом). Затем одной группе предлагается форма А теста, а другой – форма В. Через несколько дней проводится второе тестирование, причем задания меняются по подгруппам. Результаты сравниваются с помощью расчета коэффициента корреляции (по формуле Пирсона).

Метод расщепления, или метод деления теста на части, возник как результат развития метода параллельного тестирования и основывается на предположении о параллельности не только форм теста, но и отдельных заданий внутри него. Этот метод позволяет оценить надежность теста при однократном тестировании. Важнейшим условием объективности данного метода является деление теста на равноценные части. Разделить задачи теста на две части можно несколькими способами: по принципу «чет-нечет», на основе одинаковых коэффициентов сложности и дискриминативности, в соответствии со временем решения каждой задачи.

При равном количестве заданий и одинаковом их качестве первый метод при тестировании индивидуальной успеваемости, как правило, имеет коэффициент надежности чуть выше 0,9, метод параллельных форм – 0,8.

Источники неудовлетворительной надежности:

- 1) угадывание (оно в основном влияет на задания типа «истинно – ложно»);
- 2) формулировка заданий (использование понятных недвусмысленных заданий повышает надежность тестов);
- 3) величина теста (чем длиннее тест, тем он надежнее);

4) инструкции к тесту (неоднозначные инструкции приводят к снижению надежности).

Следующим важным методическим критерием качества теста является достоверность, или *валидность*. Тест называется валидным, если он измеряет то, что требуется измерить. А как мы узнаем, что тест измеряет то, для чего он предназначен? Чтобы ответить на этот вопрос, требуется определить критерий. В зависимости от вида критерия в методике тестирования различают несколько видов валидности: конкурентную, критериальную, содержательную, конструктивную.

Конкурентная валидность рассчитывается по корреляции результатов выполнения данного теста с результатами других тестов, которые доказательно предназначены для измерения того же качества или свойства. Так, если мы пытаемся установить конкурентную валидность некоторого теста, мы будем изучать его корреляцию с другими тестами, валидность которых установлена. Конкурентная валидность также полезна для установления того, что же не измеряет тест. Тест не должен иметь корреляции с тестами, измеряющими совершенно другие переменные. Конкурентная валидность не является достаточно полным аспектом валидности. Чтобы принять тест как валидный, кроме определения конкурентной валидности, требуется исследование других ее типов.

Критериальная валидность помогает установить, в какой мере результаты тестирования соответствуют данным, полученным другими способами, т.е. валидизация теста состоит в вычислении коэффициента корреляции тестового результата с внешним критерием. Основная трудность при такой валидизации теста состоит в выборе значимого внешнего критерия. На практике в качестве критерия валидности обычно используются:

- социально-демографические и биографические данные (стаж, образование, поступление или непоступление в учебное заведение и т.д.);
- производственные, научные, учебные показатели эффективности деятельности (число публикаций, оценки академической успеваемости);

- экспертные оценки выраженности черт личности, успеваемости (методы коллективной оценки, средневзвешенной оценки, ранжирования и парного сравнения);

- результаты выполнения стандартизированных методик.

Природа изучасмого явления определяет и широту используемых критериев. Более узкие критерии применяются при обосновании тестов на измерение уровня развития конкретных знаний, умений, навыков. Чем сложнее измеряемое явление, тем шире и обобщеннее должен быть критерий валидности.

Содержательная валидность, по существу, показывает отражают ли задания теста все аспекты исследуемого параметра или нет. Что это значит? Тест не является валидным вообще, он может быть валидным лишь в определенном смысле. Математический тест, предназначенный для шестого класса, не является валидным для другой ступени обучения.

Содержательная валидность определяется на основе экспертных методов. Эксперты оценивают задачи теста по шкале близости к реальным задачам, встающим перед человеком в процессе обучения, и целям предполагаемого тестирования, т.е. устанавливают связи между тематическим описанием учебного материала в учебной программе предмета и заданиями теста. Для такого сравнения можно создать матрицу соответствия учебных элементов и заданий тестов.

Конструктивная валидность непосредственно для тестов достижений не используется. Здесь речь идет о теоретическом конструкте, посредством которого результаты исследования могут быть проверены на валидность. Если эмпирические данные подтверждают гипотезу, то тем самым подтверждается психологическая концепция, положенная в основу теста, и способность теста служить инструментом измерения данного конструкта.

Итак, валидность теста устанавливается на основе серии результатов. В исследовании валидности все зависит от знаний и проницательности разработчика тестов. Она не может быть представлена в виде одного коэффициента, а определяется в результате проведения различных процедур установки валидности.

Следующий критерий качества теста – *объективность*. Мы говорим об объективности измерения в том случае, если его результаты максимально независимы от исследователя, иными сло-

вами, если различные исследователи при измерении одного и того же признака приходят к одинаковому результату. Для того чтобы получить объективную оценку знаний, все учащиеся должны быть подвергнуты одному и тому же испытанию в аналогичных условиях. Для этого унифицируется задание, время обработки заданий, пояснение к заданиям, допустимые вспомогательные средства и т.д. Также для получения объективных данных измерения в тесте необходимо использовать только те типы заданий, результаты выполнения которых могут быть обработаны объективно (в основном это относится к заданиям открытого типа).

Еще одним критерием эффективности тестов является *дискриминативность*. Она определяется как способность отделять испытуемых с высоким общим баллом по тесту от тех, кто получил низкий балл. Например, если с заданием не справляются самые сильные и самые слабые ученики, а учащиеся со средним индивидуальным баллом по тесту справляются, очевидно, что задание составлено не вполне понятно или допускается двусмысленность, о которой подавляющее большинство не знает.

Для оценки дискриминативности чаще всего применяется индекс дискриминации D . При его вычислении выделяют крайние группы, т.е. учитывают результаты наиболее и наименее успешно справившихся с заданиями учеников.

$$D = \frac{N_{p, \text{верх}}}{N_{\text{верх}}} - \frac{N_{p, \text{низ}}}{N_{\text{низ}}},$$

где N_p – количество справившихся с заданием испытуемых; N – общее количество испытуемых в крайних группах.

Существует несколько подходов к выбору крайних групп (например, количество испытуемых в крайних группах – по 27% от общего числа испытуемых).

Теперь, имея эффективный и качественный тест, педагог должен определить, как осуществлять оценивание полученных результатов. В педагогической диагностике существуют определенные нормы оценивания. Даже если нам известно, что кто-то допустил в тесте, состоящем из 30 заданий, 6 ошибок, то эта информация может интерпретироваться лишь в том случае, если мы знаем уровень сложности теста. Кроме того, мы не знаем, как

справились с тестом другие учащиеся или какое количество ошибок считается допустимым для того, чтобы признать учебную цель достигнутой. Иными словами, требуется провести сопоставительный анализ.

Результат индивидуальной успеваемости учащегося мы должны сравнить:

1) с результатами других учащихся, т.е. с социальной соотносительной нормой. Это некое согласование (с обществом, педагогам, обучаемым);

2) с прежними результатами того же учащегося, т.е. с индивидуальной соотносительной нормой. Здесь может происходить некое приращение оценки (например, ученик имел слабые знания, затем стал учиться лучше и ему в качестве поощрения ставится повышенная оценка);

3) с поставленными учебными целями (критериями), т.е. с предметной соотносительной нормой. Критерий оценки – объем усвоенного материала, выраженный в процентах.

Для оценивания результата тестирования необходимо знать целевое назначение теста. В зависимости от цели различают следующие виды тестов:

1. Нормативно ориентированные тесты. Данный вид тестирования рассчитан на то, что большинство учащихся справится только с половиной всех заданий теста за отведенное время, а меньшая часть испытуемых выполнит количество заданий больше или меньше половины. Главная задача этого классического вида тестирования – максимально дифференцировать группу учащихся по их знаниям. Если при тестировании с некоторыми заданиями никто из учащихся не справился или оказались задания, с которыми справились все, то эти задания не являются валидными. Аттестация учащихся ориентирована именно на этот вид тестов.

2. Критериально ориентированные тесты, используемые для селекции с жестким конкурсом. Например, на инженерно-экономический факультет подано 3000 заявлений, а бюджетных мест только 300. Перед проведением вступительных экзаменов проводится тестирование данного типа, и часть абитуриентов отсеивается. Уровень сложности тестовых заданий высокий.

3. Критериально ориентированные тесты с селекцией при низком конкурсе (платное обучение). Этот же тип тестов используется при контроле выполнения требований государственных стандартов. Здесь достаточно низкие требования.

В подтверждение сказанного представим классификацию обученности, предложенную А.О. Татуром и Б. У. Родионовым.

Знания классифицируются по следующим уровням:

- М – мировоззренческий минимум. Это те знания по предмету, которые должны остаться в памяти любого обучающегося. Данный уровень включает простейшие сущности, определяемые через их качества (свойства), и простейшие модели (схемы) конкретной предметной области, необходимые каждому человеку как гражданину. Преодолевая распространенное искушение дидактов постоянно увеличивать объем экстенсивных знаний, в мировоззренческий уровень мы не будем включать информацию, без которой обучающийся сможет в дальнейшем обойтись.

- Б – базовые знания, дополняющие минимум, позволяющие успешно овладеть предметом на текущей и последующей стадиях обучения. И здесь мы должны исходить из принципа минимальной достаточности.

- П – программные знания сверх базового уровня, включающие спектр основных современных знаний, в том числе основные модели объектов и явлений данной предметной области. Как отличить знания программного уровня от сверхпрограммных? Попробуйте предложить сомнительное с точки зрения дидактической классификации задание в качестве контрольного вопроса по данному предмету широкому кругу обучающихся, которые стоят на следующей ближайшей ступени обученности (например, выпускникам средней школы или вуза), и посмотрите, каков процент не справившихся с заданием (процент не справившихся показывает трудность задания). Если трудность превысит 80%, то вывод ясен: можно успешно освоить программу (перейти в следующий класс, на следующую образовательную ступень), не зная того, что проверяет данное задание.

- С – сверхпрограммные знания, рекомендованные как дополнение к программе для самых сильных обучающихся, включающие описания объектов и явлений данной области, которые необходимы будущему специалисту.

Рассмотрим умения. Как и знания, умения качественно различны. Во-первых, их можно разделить на две группы: к репродуктивным относятся умения уровней «узнавание» (I) и «воспроизведение» (II); к продуктивным, преимущественно требующим поиска самостоятельных решений, – умения уровней «умение» (III) и «трансформация» (IV). Во-вторых, умения в зависимости от знаний, на которых они базируются, отличаются в пределах каждого уровня по интенсивности.

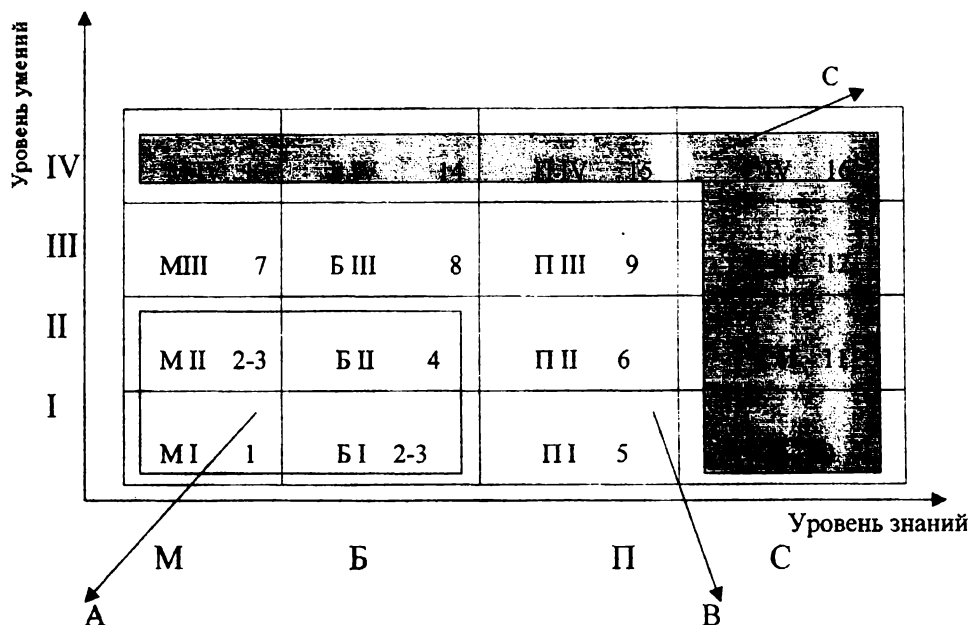


Рис. 1. Матрица обученности

Матрица обученности представлена на рис. 1. Цифры указывают порядок нарастания сложности помещаемого в ячейках учебного материала. Сложность учебных заданий обусловлена количеством и качеством операций, которые совершает учащийся при выполнении этих заданий.

Применяя данную классификацию обученности к классификации тестов по целевому назначению, мы видим, что нормативно ориентированные тесты включают задания группы Б, критериально ориентированные тесты с жестким конкурсом – задания группы С, с низким конкурсом – задания группы А.

Несколько слов хочется сказать о требованиях государственных стандартов. Они представляют собой некие согласованные социальные нормы. Контроль осуществляется с помощью «договора» составителей контрольных заданий и педагогов. (Предмет «договора» – возможное количество неуспевающих, например 15% для всей области.)

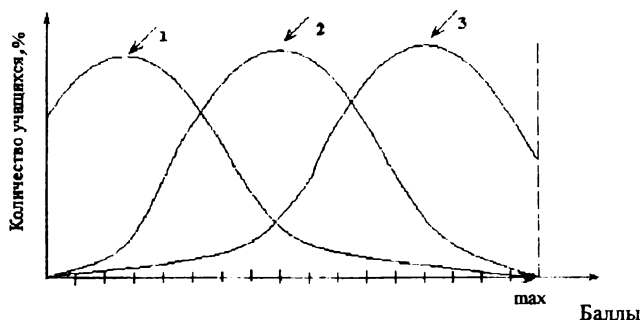


Рис. 2. Распределение тестовых баллов

На рис. 2 представлено распределение тестовых баллов. Площадь под кривой – процентное количество детей. В распределении типа 1 большинство детей набрали мало баллов (сложный тест), в распределении типа 2 – среднее количество баллов (средний по сложности тест). В распределении типа 3 большинство детей набрали много баллов (простой тест, применяемый для контроля выполнения требований стандарта). Здесь наблюдается сложность с выставлением положительных оценок.

Аттестация учащихся ориентируется на 2-й график. Тесты, имеющие такое распределение баллов, наиболее разработаны с технологической точки зрения. Рассмотрим данный график более подробно (рис.3). Он носит название нормальной кривой. Линия, идущая от высшей точки нормальной кривой вертикально вниз к точке 0 шкалы измерений, обозначает среднюю величину. Стандартное отклонение S , находящееся левее средней величины до $-4S$ и правее средней величины до $+4S$, есть величина интервала, которую получают, высчитывая расстояние между отдельными показателями и средней величиной. На рис.3 отмечено, какой

процент результатов наблюдений при нормальном распределении обусловлен стандартными отклонениями.

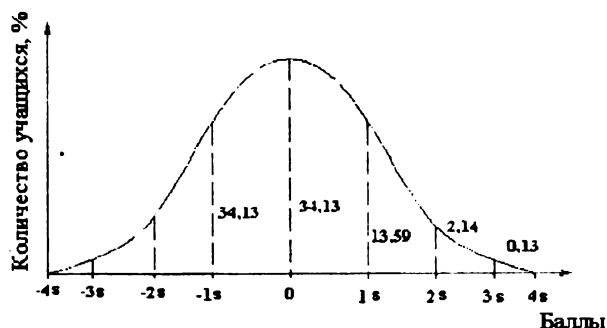


Рис.3. Нормальное распределение случайной величины

Возвращаясь к тестам, подчеркнем, что информативность тестовых норм в значительной мере зависит от того, насколько велика была выборка, была ли она стандартизована или нормирована и каким образом составлялась. При тщательно составленной выборке рассчитывают на то, что она будет отражением соответствующей генеральной совокупности, для которой должны действовать эти нормы. Поэтому автор теста должен подробно описать в приложении к тесту составление и величину выборки. Применительно к школьным тестам шкала норм может быть признана достаточно информативной лишь в том случае, если состав выборки по городам, сельской местности, полу учащихся и т.п. пропорционален соответствующей генеральной выборке и отбор перечисленных групп был случайным (не преднамеренным).

Покажем возможный вариант оценивания результатов тестирования при использовании нормативно ориентированных тестов (рис.4).

Важно, чтобы оценка «3» на графике размещалась правее границы угадывания Γ_{ye} . Что понимают под границей угадывания и как ее рассчитать? Случайное угадывание правильного ответа снижает эффективность теста, а также его надежность и валидность. Проблема угадывания ответов в основном сказывается при альтернативных заданиях.

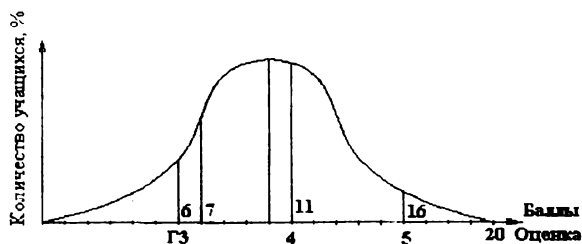


Рис.4. Соответствие оценок и тестовых баллов для нормативно ориентированных тестов

Если тесты предназначены для проверки знаний в качестве итогового контроля, то испытуемые имеют сильные мотивы для угадывания ответов. Кроме того, угадывание происходит и в случае, если требуется высокая скорость выполнения тестов в сжатые сроки. При использовании тестов с достаточно большим количеством заданий, содержащих много вариантов выбора, случайным угадыванием можно пренебречь. Однако здесь возникает еще одна проблема – оценки пропущенных заданий, на которые испытуемые не дали ответа вообще. В таком случае следует, как советует П. Клайн, ввести в инструкцию требование угадывать ответы; перед подсчетом показателей по тестам ответить на все пропущенные испытуемым вопросы произвольным образом, что будет эквивалентно случайному угадыванию. При выполнении этих требований все показатели будут искажены случайным угадыванием. Следовательно, валидность теста не будет очень сильно искажена.

Какова вероятность угадывания правильного ответа? Ответ на этот вопрос вытекает из теории вероятности. Например, если $N=30$ – количество заданий в тесте, $m=5$ – количество вариантов ответов на каждое задание, то вероятность угадывания правильного ответа одного задания $p = \frac{1}{m} = \frac{1}{5}$. Тогда среднее количество правильно угаданных ответов по всему тесту $N_{\text{ср}} = \frac{N}{m} = \frac{30}{5} = 6$.

Определим границу угадывания $\Gamma_{\text{уэ}}$ по следующей формуле:

$$\Gamma_{\text{уэ}} = \frac{3}{2} \cdot \frac{N}{m} = \frac{3}{2} \cdot \frac{30}{5} = 9.$$

Данный подход обеспечивает погрешность не более 10% (рис.5). Погрешность определяется, когда ведется длительное наблюдение.



Рис.5. Граница угадывания правильных ответов

Итак, в данной статье мы затронули обобщенные теоретические основы технологии изготовления тестов. Тест достижений, являющийся измерителем качества усвоенных знаний, как и всякий измерительный инструмент, имеет ряд свойств:

- 1) надежность измерения;
- 2) валидность (дискриминативность);
- 3) целевое назначение измерения;
- 4) погрешность измерения.

Вопросы практического составления тестов в рамках данной статьи не рассматривались.

В заключение отметим перспективы развития технологии тестового контроля. Многие преподаватели все еще считают тесты просто одной из многих форм контроля параметров человека и спрашивают: «Зачем нужно изобретать какие-то новые формы контроля, если существует масса традиционных методик?» Мы считаем, что, с одной стороны, профессионально выполненные тесты – наиболее объективная, дидактически строго выверенная, психологически максимально комфортная, технически предельно простая и экономичная методика измерения различных качеств человека. С другой стороны, современная тестовая технология принципиально отличается от традиционных недостаточно качественных оценок дотестовой эпохи образования благодаря введению в педагогику строгих математических методов оценивания учебных достижений как отдельной личности, так и любых коллективов; как отдельных специальностей, так и целых отраслей образования.